AFRL-IF-RS-TR-2006-243
Final Technical Report
July 2006

PHRASE-BASED MULTIMEDIA INFORMATION EXTRACTION

StreamSage, Inc.

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2006-243 has been reviewed and is approved for publication

APPROVED: /s/
SHARON WALTER
Project Engineer

FOR THE DIRECTOR: /s/

JOSEPH CAMERA, Chief
Information & Intelligence Exploitation Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
JULY 2006	Final	Dec 03 – Dec 05

4. TITLE AND SUBTITLE

PHRASE-BASED MULTIMEDIA INFORMATION EXTRACTION

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA8750-04-2-0015

5c. PROGRAM ELEMENT NUMBER
62805F

6. AUTHOR(S)

Eric Cohen and Evelyne Tzoukermann

5d. PROJECT NUMBER
459E

5e. TASK NUMBER
4D

5f. WORK UNIT NUMBER
U1

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
StreamSage, Inc.
1202 Delafield Place, NW
Washington DC 20011-4418

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFRL/IFED
525 Brooks Rd
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2006-243

12. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA# 06-498

13. SUPPLEMENTARY NOTES

14. ABSTRACT
StreamSage proposed to develop a prototype software system that would specifically deal with the two primary challenges of speech data on the performance of information extraction: degraded input data and the time-based nature of the content. In order to overcome these two challenges, this effort focused on two general areas:mitigating the degraded quality of speech data and improving entity identification. Technologies developed under this project for audio/video named entity identification and end user access to relevant information could have tremendous value for both military and commercial entities.

15. SUBJECT TERMS
Information extraction, audio extraction, named entities, speech recognition, topic segmentation, topic identification

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Sharon M. Walter
U	U	U	UL	26	19b. TELEPONE NUMBER (*Include area code*) N/A

TABLE OF CONTENTS

LIST OF FIGURES

1. Introduction

1.1 List of Contributors

The following have contributed to this project: Evelyne Tzoukermann, Tim Sibley, Anand Kumar, Robert Rubinoff, Shachi Dave, Abby Elbow, David Houghton, Goldee Udani, Dave Dulansey, Hemali Majithia, Padget Rice, Tony Davis, and Phil Rennert.

1.2 Background

Since the early 1990's, technology has revolutionized information distribution as millions of individuals and organizations have been able to distribute text content quickly and cheaply. The resulting deluge of information has been managed, in part, by advanced search, personalization, and knowledge management applications that allow users to quickly find and manipulate relevant content.

Over the next ten years, a similar revolution will occur as the network infrastructure becomes equipped to distribute audio/video content. Intelligence analysts are at the forefront of this audio/video content distribution revolution. Today, intelligence analysts already have access to thousands of hours of potentially relevant audio/video content from radio and TV broadcasts, transmission intercepts, surveillance cameras, meetings, and other sources.

In other fields, although most commercial enterprises are not as advanced as the intelligence community in their use of audio/video content, many organizations are starting to produce large volumes of audio/video material. Financial enterprises, educational institutions, civilian government agencies, and radio/TV broadcasters have been producing such large quantities of content that they now need more advanced tools for identifying named entities (such as companies or executives in the financial markets) or improving user interfaces for navigating within content (for instance, students searching through thousands of hours of medical school lectures).

While the intelligence agencies are resolving the short-term issue of developing sufficient digitization, storage, and network infrastructure to capture and distribute large volumes of audio/video content, the more complex challenge of improving analysts' ability to extract information from this content remains unsolved. This project focused on improving analysts' ability to extract information for audio/video content that contains spoken language.

Speech is one of the richest and most accessible information sources in audio/video content, but presents two primary challenges to effective information extraction:

- Speech recognition (SR) output is seriously degraded, lacking punctuation and capitalization, having significant word error rates, and mirroring the disfluencies of spoken language. These factors greatly reduce the performance of standard text information extraction techniques.
- Listening to speech can occur only linearly and not much faster than real-time. Consequently, audio/video content requires much more sophisticated user tools and interface than does text, for identifying relevant content and navigating within the content.

1.3 Project Directions

In order to begin addressing these information extraction challenges, this project undertook to combine existing audio/video analysis tools with technologies developed under the tasks of this project.

StreamSage proposed to develop a prototype system that would specifically deal with the two primary challenges of speech data on the performance of information extraction: degraded input data and the time-based nature of the content. In order to overcome these two challenges, StreamSage effort focused on two general areas:

- **Mitigating the degraded quality of speech data**: StreamSage was to combine existing proprietary technologies with new modules developed under this project to extract linguistic information from degraded SR output, and use new analyses of this information to improve the word accuracy of the SR system, and to customize the vocabulary of the SR system.
- **Improving entity identification**: StreamSage was to improve techniques for identifying named entities and other content-central entities in text and SR output. This was to improve the data available for entity tracking and identifying content and entities of interest, and to improve SR results on out-of-vocabulary entities.

Technologies developed under this project for audio/video named entity identification and end user access to relevant information could have tremendous value for both military and commercial entities. Within the intelligence community, analysts are tasked with extracting and analyzing information from thousands of potential sources. Technologies pursued in this effort could tremendously facilitate this process for the analysts, saving time, money, and potentially lives.

1.4 Project Overview

The goal of this project was to improve speech recognition (SR) technologies, and to build a prototype system capable of:

- accurately identifying named entities within audio/video content;
- accurately identifying the portions of audio/video content related to topics or named entities of interest; and
- providing end users with an intuitive user interface for accessing the information in the audio/video content.

The project was planned to conduct research in several areas that would make this final system possible:

- New approaches to improved speech-recognition (SR), including two-pass SR.
- Improvements in named-entity recognition in SR output.

1.5 Results

Overall, the demonstration platform delivered at the conclusion of the project shows the feasibility of this development, and progress towards its various goals.

The demonstration platform, as delivered:

- Allows a user to choose named entities of interest, and store that list of interests in a personal profile.
- Presents a list of audio files corresponding to the named entities or topics in the user’s profile.
- Calculates and displays the relevance of the displayed segments.

Project research has:

- developed techniques showing substantial SR improvement, particularly on named entities;

- developed techniques in topic segmentation and topic identification, making possible the improvements to SR;
- improved technology for named entity identification.

2. Underlying Technologies

2.1 Indexer and Relevance Intervals

As a foundation for many of this project's tasks, development has built on existing StreamSage technologies for indexing audio/visual content. The central StreamSage indexing platform performs these tasks:

1. Accepts as input a digitized audio/visual file.
2. Performs SR to create a transcript with time-synchronization information.
3. Lemmatizes words in the SR output to root forms, as well as expanding contractions and performing other basic pre-processing of the SR output.
4. Performs basic part-of-speech tagging of the SR output.
5. Analyzes the transcript for *content words* (nouns and multi-word units, excluding stop words).
6. Detects and resolves certain types of anaphoric references.
7. Compiles co-occurrence data (described below) for these content words.
8. Calculates centrality values for content words and resolved references. Based on syntactic information and other cues, this calculates how central each item is to the meaning of its containing sentence.
9. From the data calculated in steps 2–7, derives *relevance intervals* for every content word in the media: these are complete lists, with time references, of those intervals in the media file which are concerned with that term, together with a score of how relevant the term is.

Relevance intervals may be non-contiguous. A single term may have a relevance interval composed of many non-contiguous portions; and conversely, any instant in the source media is likely included in the relevance intervals for many different terms.

These relevance intervals are central to the topic segmentation and topic identification tasks described below.

2.2 The Co-occurring Word (COW) Model

Much of this project's research in topic segmentation, topic identification, and finding custom content for SR training, depends on and extends StreamSage's co-occurring word model (the COW model). This is a large-scale model of word co-occurrence patterns, a database built on the analysis of over 350 million words of English text from a corpus of six years of *New York Times* articles, as well as from further co-occurrence data derived from specialized domains. This model measures the frequency with which content words co-occur, within a window of 50 words to either side, over the entire corpus.

From this data, the model obtains the COW value of any two words in the corpus — this is a normalized measure of the mutual information between the two words, or how much more (or less) those two words co-occur than they would by chance.

The model is refined by chunking major constituents, so that words with a higher centrality receive greater weight. It has also been extended to multi-word units, many of which are idiomatic or otherwise noncompositional (e.g. “fly ball”, “kick the bucket”.)

3. Task: Topic Segmentation

Much of the utility of this project, and many of its other technologies, depend upon an accurate and intelligent means of automatically identifying areas of topical interest within audio/video content:

- Techniques for improved SR via topic-specific training or vocabulary require accurate detection of topic changes within heterogeneous audio/visual material.
- A platform for giving a human analyst access to intervals of interest within heterogeneous audio/visual material requires techniques for identifying those portions of interest.
- Such a platform also must begin and end excerpted segments both accurately and appropriately: including necessary context, and continuing to the end of the topic (rather than simply cutting in and out of audio/visual content at the first and last occurrences of keywords), but also omitting irrelevant material to save analyst time and concentration.

Toward those ends, this project has developed and extended several techniques for automatically detecting topic boundaries in audio/visual content. Note that these techniques are dynamic, in that they recognize topics without relying on a pre-determined set of topics or keywords.

3.1 Content-based Segmentation

Project research has focused on these content-based dynamic techniques for topic segmentation:

3.1.1 Noun-Link Segmentation Algorithm

This algorithm, in outline, tags all nouns that occur twice within a *maximum link distance* (MLD) as potential topic words, and links the sentences that contain them. Any sentence boundary spanned by no links is a potential topic segment boundary.

Refinements:

- Testing shows best segmentation results for an MLD of 12 sentences, for clean text from a manual transcript, or 20 sentences for SR output text (as SR output typically contains many sentence fragments tagged as sentences).
- A frequency filter eliminates nouns that are found frequently throughout the content (nouns such as “thing”, for instance) from the nouns used to create links and determine boundaries.
- Multi-word units (that is, fixed phrases and some compound nouns, such as “fly ball”, “White House”) are treated as single nouns for tagging and finding links.

3.1.2 Choi’s Algorithm

Choi’s algorithm is based on the principle that sentences within a topical segment will be much more semantically similar to each other than they are to sentences in other segments.

StreamSage’s implementation of this concept uses the COW model (described above) to measure semantic similarity. For any two sentences, this project measures their similarity from the COW values of the words in those sentences, normalized to account for sentence lengths. From this, the procedure can calculate and store the similarity between every pair of sentences in the content.

Finally, measuring within a window of five sentence before and after each sentence, the algorithm detects local minima in the summed similarity between each sentence and its neighbors. These minima are tagged as potential segment boundaries.

3.1.3 Induced N-grams Algorithm

The induced n-grams algorithm detects multi-word markers that occur frequently near segment boundaries. To do this, it starts with potential segment boundaries detected by other methods, such as the noun-link segmentation algorithm and Choi's algorithm, described above. Then, operating on as large a corpus as possible of material similar to the content of interest, it detects one- two- or three-word markers (unigrams, bigrams, and trigrams) found disproportionately near these potential segment boundaries.

This algorithm is strikingly successful on structured content, such as news broadcasts, where marker phrases such as "In other news . . .", or newscaster names, are very consistent markers of topic changes.

3.1.4 Combined Results

Testing shows that simply combining all the potential topic boundaries detected by the three above algorithms produces too many topic boundaries — high recall, but low precision. Rather, because the induced n-gram algorithm has high precision, any boundaries detected by that algorithm are weighted highly in combining the results. Potential boundaries detected by either the noun-link algorithm or Choi's algorithm that fall within a sentence or two of potential boundaries detected by induced n-grams are not accepted, as it is overwhelmingly likely that only one useful boundary should be placed there, and that the induced n-gram algorithm has located it correctly. Otherwise, Choi's algorithm is used as a check on a tentative noun-link boundaries, and vice versa.

3.2 Prosody-based Segmentation

All of the topic segmentation techniques described so far base their analysis on lexical information — the words identified in the audio stream. The project has also researched topic segmentation via *prosodic information:* acoustic information about the speech in the content, such as fundamental frequency, frequency changes, volume, pauses, and time-related features of phones.

We have used these prosodic cues successfully for detecting commercials in news broadcast. This, in turn, improves the precision of related technologies: correctly paring commercials from news content results in better base data and better accuracy in determining topic words (described below) for the content of interest (the news content).

4. Task: Topic Identification

Several tasks described below, including techniques for two-pass SR improvement, depend upon the ability to automatically identify the topic of a segment, and in particular to identify central *topic words* associated with this segment. Some of these topic words will be named entities of interest, others will be other nouns or noun phrases that are typical of the topic and less common in other topics. (This work depends upon successful topical segmentation of the audio/visual content, as described in the preceding section.)

4.1 Topic Word Discovery

This project's approach to finding these topic words begins with a pool of potential topic words for any segment — the *content words* in that segment. A content word is a noun phrase, or a compound headed by a noun phrase. A content word compound may be an adjective-noun compound ("potable water"), a noun-noun compound ("birthday cake"), or a multi-noun or multi-adjective extension of such a compound, ("director of the department of the interior"). This approach also maintains a list of topically general nouns, such as "everyone" and "thing" that may not be content words.

Next, given a segment's list of content words, the task is to narrow that list to those words that are most typical of the segment's topic. Research has arrived at several techniques for doing this:

- **First in segment**

  This consists of capturing potential topic words that occur early in the segment, as the topic under discussion is often identified here.

- **Low corpus frequency**

  Good potential topic words may be those words in the segment that occur infrequently in an appropriate large reference corpus — these are idiosyncratic words typical of the topic. For general audio/visual content, a general corpus, such as one we have used derived from the *New York Times,* or a corpus of conversational speech, is appropriate for comparison. For more specialized content, a more appropriate comparison corpus is required.

- **High segment frequency**

  Content words that occur frequently in the segment are also good potential topic words.

- **Cluster centers**

  This consists of using the COW model to calculate which potential topic words tend to co-occur with many other potential topic words in the same segment. This cluster of highly-interrelated content words is likely to be central to the topic.

A weighted sum of normalized scores from these algorithms is a useful measure of the overall best topic words. (As the high segment frequency algorithm picks out words with high topic frequency, and the low corpus frequency picks words with a measure similar to high inverse document frequency, this combined score is an extension of TF/IDF-based measures that are widely used for document categorization.)

4.2 Topic ID Testing

Starting from a corpus of 7.5 hours of recorded speech, StreamSage created a gold standard test bed by having human annotators determine relevance intervals for this content: intervals, and topic descriptions, were manually created for 949 distinct topics.

In preliminary testing, we ran the topic identification procedure described above on this test bed. Over 140 of the human-identified topics were also tagged as the most important topics using this automated approach.

4.2.1 Topic ID Sample Output

Figure 1, below, shows an example of typical low-quality SR output. The topic words calculated for this segment are *"Space Shuttle Columbia", omen, astronaut, NASA,* and *"brilliant astronauts".*

The is for for we're for a story's homecoming first-ever brilliant astronauts hurled into horror space shuttle courier broken up on reentry over Texas our Miles O'Brien was on the air good morning when word first came that something had gone terribly wrong she to out the morning time telling viewers said so not to see the space shuttle ancestry back turn recognize time good morning taxes take a look outside the space shuttle Columbia that awash in many fifteen minutes I had heard right on the top of the hour not gove the mosque communication not with some of the season to the news year about the house the shuttle Columbia is we have n't heard from the yet the time of landing a was supposed to be comedy about this moment nin sixteen expected landing time time when Navy said a been on the ground dinner rock and thing about that he is a there 's no real around is not likely diverted to down uh I knew in time this a very ominous candidates at the crew was most likely lost here 's what we 're seeing it is very significant ahead bowl sales multiple vindication of multiple five years there has the space shuttle 's three over Dallas Texas I remember sitting here a you know watching the fee for NASA listening to what they were talking about about...

Figure 1: SR of news segment, with topic words "Space Shuttle Columbia", omen, astronaut, NASA,"brilliant astronauts"

Note that most of these topic words are accurate, despite SR output including "mosque", "gove", "ground dinner rock", "taxes", "candidates", and the like.

4.3 Other Uses of Topic Segmentation and Topic Words

4.3.1 Topic Segmentation

StreamSage has used our topic segmentation technology in our commercial media indexer; this technology is deployed for CNN and other commercial content providers. It consistently performs well on degraded SR output.

In order to test this technology's portability to another language, we have performed preliminary testing in Arabic with encouraging results.

4.3.2 Topic Words for Content Clustering

StreamSage has also used topic words, arrived at as above, as the basis for topic-based clustering of documents and media segments.

- Applied to documents, this allows a user to find common topics and a hierarchical organization within a collection of documents, without manual indexing or categorization.
- Applied to media segments, this can allow a user to find the relevant audio-visual segments within a library; or, in an entertainment context, to automatically create channels of desired entertainment media as the media become available, without human indexing. Creating a personalized feed of Italian sports news, say, for the expatriate; with a sub-cluster of World Cup-related news.

In work outside the scope of this project, we have employed mathematical clustering techniques to the topic words associated with topical media segments. These techniques apply similarity measures to the topic words (similarity measures based on information in the COW tables described above), and use these results derived from the topic words to group media segments into clusters of similar content.

Results so far have been quite positive, showing good grouping into human-useful clusters, without manual indexing or organization.

4.4 Topic Segmentation and Identification in Conversational Speech

In October 2004, StreamSage began NSA-funded work on some goals parallel to those of this project, although directed particularly to the needs of processing conversational speech. The technology developed there has proved complementary to the technology developed in this project, and has advanced the overall target technology.

As a first task of the NSA project, we examined topic segmentation of conversational speech. Since conversational speech is generally composed of shorter and less coherent utterances than the testbed of news broadcasts used in this project, obtaining good topic words was appreciably more difficult. As well, SR of conversational speech is generally poorer than of broadcast corpora due to lower audio quality, more widely-varying vocabulary, and alternation and overlap of speakers. Nevertheless, we obtained some results pointing to the robustness of this project's technologies: Although the automatically-detected topic words obtained from conversational speech were moderately degraded compared to those obtained from broadcast news, they were still able to convey the gists of the conversations, and were therefore useful for all the tasks depending on those topic words.

5. Issues in SR Improvement

Current standard SR output exhibits several characteristics that pose serious difficulties, both for human comprehension and for subsequent automated processing.

- **Absence of punctuation**: Most crucially, this means that sentence boundaries are not indicated. Neither is speaker change, or quotation; and phrase boundaries are similarly unmarked.
- **Absence of capitalization**: For named-entity extraction in text, capitalization is by far the best single cue for identifying named entities — generally, proper nouns. SR output only capitalizes words that are capitalized in the SR system vocabulary. Thus, out-of-vocabulary proper nouns words and multi-word named entities will generally be incorrectly uncapitalized in SR output. Exacerbating this problem, named entities will be disproportionately out-of-vocabulary for any standard dictionary.
- **Speaker disfluencies**: Spoken language exhibits unique characteristics not found in written language, such as mid-utterance corrections, frequent use of phrases that are not complete sentences, filler words ("umm", "ah", "err . . ."), and the abrupt start and stop of phrases. Even perfectly transcribed conversation, lacking the prosodic and other cues of live speech, will be very difficult to understand.
- **SR errors**: Current standard SR typically exhibits word error rates of 10-50% for large-vocabulary, speaker-independent SR. Speaker-dependent SR (training an SR for one particular speaker) and limited-vocabulary SR (as in systems for navigating voice mail) can achieve much better results, but in clearly much more limited domains. For the needs of intelligence analysis, as well as many commercial technologies, the necessary task is large-vocabulary speaker-independent SR.

5.1 Increasing SR Accuracy

Previous work has suggested certain paths to improving SR results.

5.1.1 Content-appropriate vocabularies and training

An SR system is limited in its abilities by the extent of its internal dictionary. Every word in an automatic-speech-recognition transcript is drawn from this dictionary. A general-purpose dictionary, then, will be particularly inadequate to input that utilizes a specialized, technical, or otherwise unusual vocabulary. SR, in such a case, will be less accurate, generating many more entirely incorrect words, or areas of failed recognition.

For example, and as addressed in another StreamSage project, NASA has an extensive series of lectures related to astrophysics. These lectures contain technical language uncommon in every day conversation. As a result, the off-the-shelf SR software performs very poorly in accurately transcribing these lectures. the words the speakers use. However, this software can be trained with vocabulary-appropriate content: articles and other text related to astrophysics. After such training, SR on these NASA lectures is greatly improved in accuracy despite the technical language.

5.1.2 Exact-transcript training

Speech recognition software accuracy can also be improved using an exact transcript of the audio/video file. Obviously, SR is not needed for a file for which a high-quality human-generated transcript exists. But for an application in which a library of files and matching transcripts has already been created, these can be used to train an SR system for high-quality results on similar files which have not been transcribed.

Training SR software on media files with accurate transcripts significantly decreases the word error rate of the SR software on similar files. During a pilot project with CNN, we were able to train the software with transcripts to the extent that the SR results exceeded 80% accuracy.

5.1.3 Two-Pass SR

As described below, much of this project's work on improving SR results aimed at incorporating additional linguistic information via *two-pass SR*. In this technique:

The basic model for this project is based on the following workflow:

1) Input speech is recognized using an off-the-shelf speech recognizer.
2) The results of that SR are analyzed to yield additional linguistic data about the source speech. Although the first-pass SR is inaccurate and incomplete, enough content words are recognized for useful input to the techniques described in this report.
3) The results of the analysis in step 2 are used to adjust the language model used in the speech recognizer. This consists essentially of instructing the SR software that various words and word combinations are more or less likely than predicted by the unimproved first-pass language model.
4) The same SR software, but using this improved language model, starts fresh from the input speech audio.
5) The resulting SR output is improved relative to the first-pass SR baseline.

This information flow is shown in figure 2:

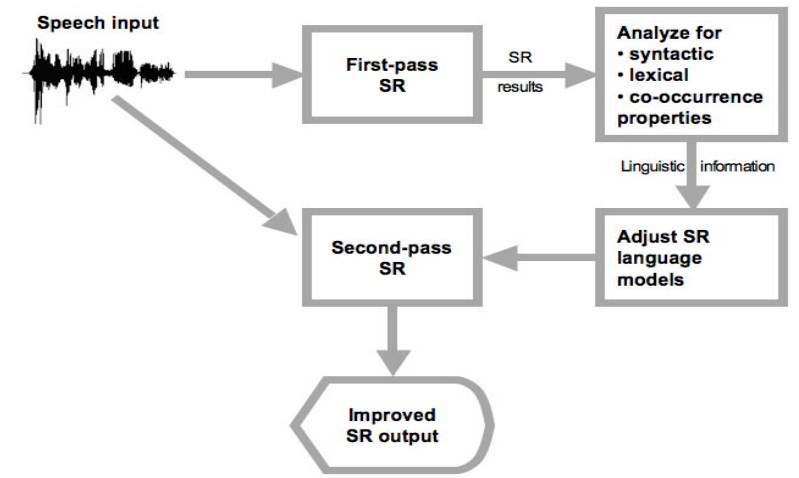


Figure 2: Information flow and order of operations in two-pass SR

6. Task: Extracting Named Entities from Speech Data

One component task of SR is recognizing named entities and other entities of interest. For SR, this is an aspect of the task of recognizing out-of-vocabulary entities, as named entities are less likely to be recognized using a standard dictionary. This poor recognition is at odds with Air Force goals, as these named entities are also more likely to be central to the information of interest to an intelligence analyst. For information retrieval, tagging and tracking entities of interest (primarily individuals, although also places, organizations, institutions, and the like) through multiple references in multiple documents and audio/visual sources, is a key task in making the relevant information automatically available to an analyst.

This project extended several technologies to advance our abilities in recognizing named entities.

6.1 Evaluation Metric

In order to evaluate techniques for named entity extraction in this project, we developed a testbed of hand-tagged data, and an appropriate evaluation metric. For evaluation, we examine the *type* and *token* measures:

- **type**: counts detected occurrences of named entities, versus all possible correct detections of named entities in the test data. Its disadvantage is that this measure penalizes a score for every missed occurrence of a single named entity (“Jane Doe”, for example) that may recur many times in the test data. Our techniques for named entity identification will tend to either detect or miss *all* occurrences of a named entity, so type can artificially decrease algorithm scores for missing a single name that is common in the test corpus.
- **token**: counts the number of named entities that are detected correctly *at least once*, versus all the different named entities in the test data. For our purposes this is a more accurate measure of what percentage of named entities an algorithm detects, as it does not cloud the numbers with repeated misses.

In evaluation, the project measured precision and recall for both the type and token measures, broken into several categories of named entity.

6.2 Identifying Named Persons

This portion of the project concentrated on improving the identification of person names in text. This included techniques for recognizing both full names and partial names in transcript text, and the much more difficult task of recognizing them in noisy, uncapitalized, SR output.

Unless otherwise noted, the project's developments in this area were based on using a finite-state machine to identify person names. This machine transitions through states corresponding to prefix/honorific, first name, middle name, last name prefix (such as "von" or "de") and last name.

- **Honorifics and name cues**: Certain phrases or titles are very good clues that a name will follow. Titles in this category include words such as "Mr.", "Mrs.", "Dr.", and so on. Words such as "named" are often also good indicators. We have worked to expand this list, while also checking results for increased recall at too high a cost in accuracy.
- **Co-reference for gender assignment**: When assigning gender to a named entity, employing more data leads to better results. Using only local features, an algorithm cannot tell whether *Smith,* in "Smith sat down", is male or female. If the content includes the earlier named entity *Jane Smith,* and the algorithm can accurately identify this as coreferential to *Smith,* information is available to tag *Smith* as female. We improved the use of co-reference data in assigning gender to named entities.

  We also improved the lists of male and female names and other local cues used to assign gender to named entities.
- **Delimiting adjacent names**: In a context such as "Maria, Tanner, and Bob went fishing", the ideal algorithm would identify *Tanner* as a separate named entity, rather than as Maria's surname. Improvements were made to the finite state machine for correctly tagging constructions of this type.
- **Other finite-state machine improvements**: The project also developed logic that allowed better identification of names with periods — *J. K. Ramirez, T. Grant Smith, Lita S. Jones;* names with commas — *Hector Jones, Jr.;* and conjoined names, such as *Sherlock and Judy Holmes.*

Using both the type and token metrics (described above), we tested these extensions and improvements to the name identification module on a small testbed taken from a print corpus. This testbed included common person names in a variety of forms and contexts. We found substantial improvements to both precision and recall using these improvements in technique.

6.3 Identifying More Complex Named Entities

This portion of the project concentrated on identifying more complex entities than person names, using techniques in addition to the finite-state machine mentioned above.

The major effort of this subtask was implementing the NOMEN machine-learning algorithm[1] to do data-mining for types of entities more complex and difficult to identify than regular proper names. Yangarber and Grishman's original application focused on identifying the names of

[1] Yangarber R., Lin W., Grishman R. "Unsupervised learning of generalized names", Proc. 19 Computational Linguistics, 2002.

diseases. These are harder to identify than person names, even in fully capitalized and punctuated text, because of the wide variety of forms represented in disease names (compare “E. Coli”, “yellow fever”, “Charcot-Marie-Tooth Disease”, “(new) variant Creutzfeldt-Jacob disease”, and “Bacillus anthracis infection”). The task we chose was similarly complex: extracting types of equipment from a corpus of technical documents in the aerospace domain.

The strengths of this algorithm are that it bootstraps from an initial set of seeds and knows when to stop learning due to competition between categories. Given a small set of positive examples (like “tuberculosis”, “E. coli”, and “yellow fever” for diseases, or “launch vehicle”, “booster rocket”, and “radar wind profiler” for aerospace equipment), the NOMEN algorithm:

1. Tabulates the most frequent n-gram contexts in which these examples occur. (These are contexts in which the initial seeds are likely; and by extension, in which similar named entities may be likely.)
2. Finds constituents which occur frequently in those contexts discovered in step 1. (These are new constituents — hopefully, named entities — which occur in the same contexts as the initial seeds.)
3. Repeats steps 1 and 2, iteratively discovering new contexts and new named entities in the corpus. Iteration ends when no new examples can be found.

Initial results with this algorithm were promising, showing success for the approach in a new content area. Although time and funding did not permit further pursuing this sub-task, we identified these approaches and questions for further investigation:

- To improve recall, larger text-rich content-appropriate corpora will be needed.
- Some testing was performed on more diverse sets of entities than aerospace equipment types. For finding more named entities and other entities of interest in a larger realm, hand-categorization of the initial seed entities into fine divisions, would be needed to improve precision.
- Both this project and the literature to date used only n-gram contexts in step 1, above. It may be fruitful to enrich the patterns that NOMEN learns by incorporating more syntactic information (e.g., using part-of-speech tagging to approximate shallow parsing) and/or semantic information (e.g. the semantic categories from the *Cambridge International Dictionary of English*, or a content-appropriate ontology).

7. Task: Improved SR Through Topic-Specific Training

This aspect of the project aimed at improving SR through focused training of the SR system on materials topically similar to the content of interest.

- This should greatly improve recognition of out-of-vocabulary items, particularly named entities, as the SR system trains specifically to make these items part of its vocabulary.
- Given topic segmentation (as described above), this allows an SR system to adjust its models to the differing vocabularies of a heterogeneous audiovisual file.
- The system accomplishes this re-training entirely automatically, without need for a human editor to identify topics or provide appropriate corpora for vocabulary-appropriate training.

7.1 Automated Retrieval of Topic-Appropriate Training Material

Topic-specific training is of no use if human intervention is needed to either identify the topic of an audio segment, or to retrieve similar material for training. The topic words, however, provide cues to automatically finding similar training material. For this project, we developed techniques to retrieve this training material from the most general and easily available source — the World Wide Web.

In particular, we developed techniques for automatically retrieving material from various newspaper Web sites, and through the Google search engine. In order to pull article content off various newspaper sites, we developed custom scrapers, programs to extract information from Web pages. This was done for several large news sites: the *Wall Street Journal*, the *New York Times*, the *Washington Post*, and Google News. These scrapers automatically retrieve and process Web pages. They find the content, remove formatting and other HTML and CSS markup, remove boilerplate text present on every page, and remove advertisements and extraneous links. The resulting content — the text of the articles — is then in a form suitable for creating domain-specific vocabularies.

7.2 Topic-Specific Training Using Topic ID Information

Once we had developed the retrieval technology described above, the project built and tested a system with the following processing steps:

1. First-pass SR of the material of interest, using a standard n-gram model for SR. This provides enough recognized content for the next step.
2. Topic segmentation.
3. Finding topic words for the topical segments — topic identification.
4. Using the topic words to find similar content via Web search engines for each topical segment.
5. Re-training the SR engine using this new, topically appropriate vocabulary.
6. Second-pass SR of the material of interest, using the new topic-appropriate language model.

A use-case can illustrate this technique. “General John Abizaid” will likely be out-of-vocabulary for an SR language model based on general English-language content. However, for military news mentioning this general, even low-quality first-pass SR will yield topic words (such as “military”, “United States”, “Iraq”, and so on). Web search on those topic words will yield many Web pages that also mention General John Abizaid. After training on this content, second-pass SR will successfully recognize any mentions of General Abizaid in the content of interest.

7.2.1 Testing and Results

This task used a test bed of 63 broadcast television recordings, of widely varying news, entertainment, and sports content. Each broadcast was treated as single-topic from the nature of the source media. This varied testbed included a high proportion of named entities (news, sports, and entertainment figures), most of them out-of-vocabulary.

For each media file, ~100,000 words of English text were fetched from the Web via the Google search engine, operating on the topic words of that file. The source text was pre-processed to remove formatting, repetitions, and other bad text, yielding the 100,000 words of content for language model training.

In this test, recall on entity recognition improved from 50% with the base n-gram model to 61.2% with the topic adaptation, a 22.4% reduction in word error rate on entities.

Even more promising, these accuracy gains were seen consistently across transcript accuracy levels. For example, for those files for which the base SR language model produced greater than 80% recognition accuracy on entities, the topic adaptation decreased the recognition error rate on entities by 25.3%.

7.3 Topic-Specific Training Using File Metadata Information

In parallel with the effort just described, we also investigated another approach to finding appropriate topic specific training material. In this approach, we tested the technique of finding related Web material using information in the audio file's *metadata* — information associated with the file, such as its title, source, keywords, and any other useful data that might be available about an audio file.

The project therefore developed procedures to automate the process of extracting information from file metadata. We wrote Perl scripts to process hand-generated audio file metadata (this metadata was hand-generated only for easy preliminary testing), and to automatically create lists of search items for Web queries from this metadata.

The rest of this testing was similar to the procedures just described: for each file, retrieve relevant training material using Web search engines. Use this training material to re-train the SR engine. Finally, perform second-pass SR, using this re-trained language model, and compare the output to first-pass SR results.

Using this approach, the language models generated from automatically collected text improved SR recall of high-value terms by an average of 6.7%, with a standard deviation of 8.4%.

8. Task: Improved SR Through Syntactic Models

Under this project, we began the task of improving SR through the use of syntactic and lexical information not available to traditional trigram-based SR. In particular, we began extending the COW model data on the co-occurrence of words to include information on additional language features, such as syntactic and lexical information. Our current effort, begun as part of this project and continuing now with other funding, gathers statistical large-corpus information about:

I. Co-occurrence of words or multi-word units (MWUs) with other words (or multi-word units). This is the existing COW model described above.

II. Co-occurrence of words (or MWUs) with words (or MWUs) in a particular syntactic role. For instance in "The dog is chasing the cat", the model counts the co-occurrence of *dog-as-subject* with *cat-as-object*.

III. Co-occurrence of words (or MWUs) with CIDE semantic domain codes.* For instance in "The dog sat in the armchair", the model counts the co-occurrence of *dog* with CIDE code 805, with code 194, and so on.

* CIDE, the *Cambridge Dictionary of International English*, places every noun in a tree of about 2,000 semantic domain codes. For instance, "armchair" has the code 805 (Chairs and Seats), a subcode of 194 (Furniture and Fittings), which is a subcode of 66 (Buildings), which is a subcode of 43 (Building and Civil Engineering), which is a subcode of 1 (everything).

IV. Co-occurrence of words (or MWUs) in particular syntactic relations. For instance in "The dog is chasing the cat in the road", the model counts the co-occurrence of *chase-as-verb* with *dog-as-subject*.

V. The extension of IV, above, to semantic roles of nouns. Here the model counts the co-occurrence of *chase-as-verb* with *ANIMATE-as-subject*.

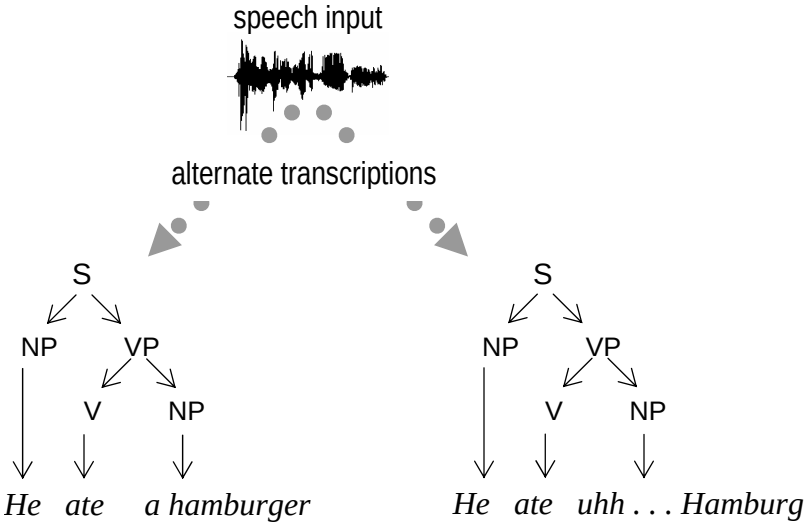
VI. The extension of V, above, to syntactic properties of verbs. Here the model counts the co-occurrence of *TRANSITIVE-verb* with *dog-as-subject* and with *ANIMATE-as-subject*.

8.1 Relationships for SR Improvement

This work is grounded in the earlier COW model which has been so useful to key StreamSage technologies in audio indexing, topic segmentation, and topic identification. Unfortunately, the COW model does not work well for SR correction, because it measures how often words co-occur within a relatively large window. Within that window, only some of the co-occurring words will be strongly correlated; indeed, some will be only very weakly correlated. Thus, the COW values of a candidate transcription with other words in its window are not a good measure for choosing among competing transcriptions, or for detecting SR errors. Preliminary experiments in 2001 using the COW model to identify possible SR errors identified at least as many false positives as actual errors.

Unlike co-occurrence in the COW model, co-occurrence within specific syntactic relations provides much more detailed and much more short-distance information about word correlations, exactly the sort of information the we hope will serve to inform better automated SR choices between candidate transcriptions.

Consider the following problem in speech recognition:



The audio data is equivocal between the two candidate transcriptions shown. Syntactic correlation data, however, can show that "hamburger" (both a Food and a Substance, according to *CIDE*) is a much better candidate for the object of "eat" than "Hamburg" (a Place).

Note that the generalized data collected (information on Food, Substance, and Place) makes this information available, even if the precise co-occurrence data ("he ate a hamburger") is sparse in the corpus.

8.2 Parsing for Relationships

An essential first step in obtaining the data necessary for this extension of a co-occurrence model is parsing a large corpus in order to be able to detect the syntactic relations of interest. To do this, StreamSage obtained a license to the Xerox Linguistics Environment (XLE) parser[2] from PARC (formerly the Xerox Palo Alto Research Center). After an extensive survey, we identified this parser as the best tool for this use. This tool, based closely on the lexical functional grammar (LFG) formalism, is the first used at StreamSage to provide parsing data of English text both reliable and fast enough to be used for SR language models. An additional advantage of XLE over many other parsers is its robustness under difficult inputs: it provides parse trees of ungrammatical constructs and provides probability estimates for the constituents it is able to form. This ability is crucial to parsing the output from SR systems, which will necessarily contain many gaps and errors.

At right, figure 3 shows an XLE parse of a sample sentence. The parser provides information about syntactic structures and lexical roles, both of which are used in constructing the co-occurrence tables. This example sentence is unambiguous and well-formed. In more complex cases, admitting multiple parse trees, XLE output describes the multiple possible syntactic and lexical structures of the input.

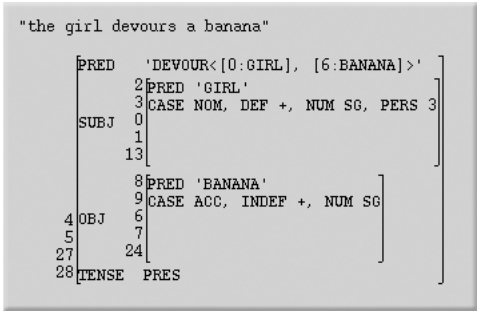


Figure 3: Sample parse from the XLE parser of an unambiguous, well-formed sentence

Although the task of parsing a large corpus was begun as part of this project, it was continued following the end of project funding. Construction of large-scale databases based on a corpus of *New York Times* articles totaling ~325 million words, including co-occurrences of types I–VI, above, was completed in April, 2006. This effort parsed all sentences from the corpus having a sentence length of 25 or fewer words. (Parsing longer sentences was impractically slow and increasingly inaccurate.) This totaled about half the word count (60% of the sentences) of the original corpus.

Testing has now begun towards using this data for SR improvement and related tasks.

9. Task: Analyst's Platform

As the final task of this project, we constructed a demonstration "analysts' platform" showing these technologies ready for use, and tied together into an application aimed at Air Force goals.

9.1 Use Case

The use case envisions a large pool of digitized audio/visual media. This content pool is always updating, heterogeneous both within and among files, and available to a group of intelligence analysts. Each analyst has his/her own areas of interest. To save time and effort, each analyst

[2] http://www2.parc.com/istl/groups/nltt/xle/ ; http://www2.parc.com/istl/groups/nltt/xle/doc/xle_toc.html .

wants to know when new content is added that may fall within their areas of interest; and to be able to quickly find and view those files *or portions of files* that are of interest. The platform should store a profile of each analyst to make this possible, and allow analysts to change their own profiles.

9.2 Platform as Built

The demonstration platform was constructed so as to demonstrate the core capabilities of such a system, while not requiring more infrastructure development than needed for this project.

- Testbed of media files: existing corpus of NPR news stories. (Heterogeneous both within and among files.)
- Interface: interaction via Web browser, media playback via Windows Media Player.
- Analysts and profiles: stores and allows modification of single profile. (Adding the capability for multiple user accounts is an obvious extension of the technology.)
- Each profile is an *entity list* — a set of entities of interest. Each entity may or may not be present in portions of media files.

This platform, including all the files necessary to run the application and use analyst's Web front end, were delivered to the Air Force at the conclusion of this project.

9.3 Using the Platform

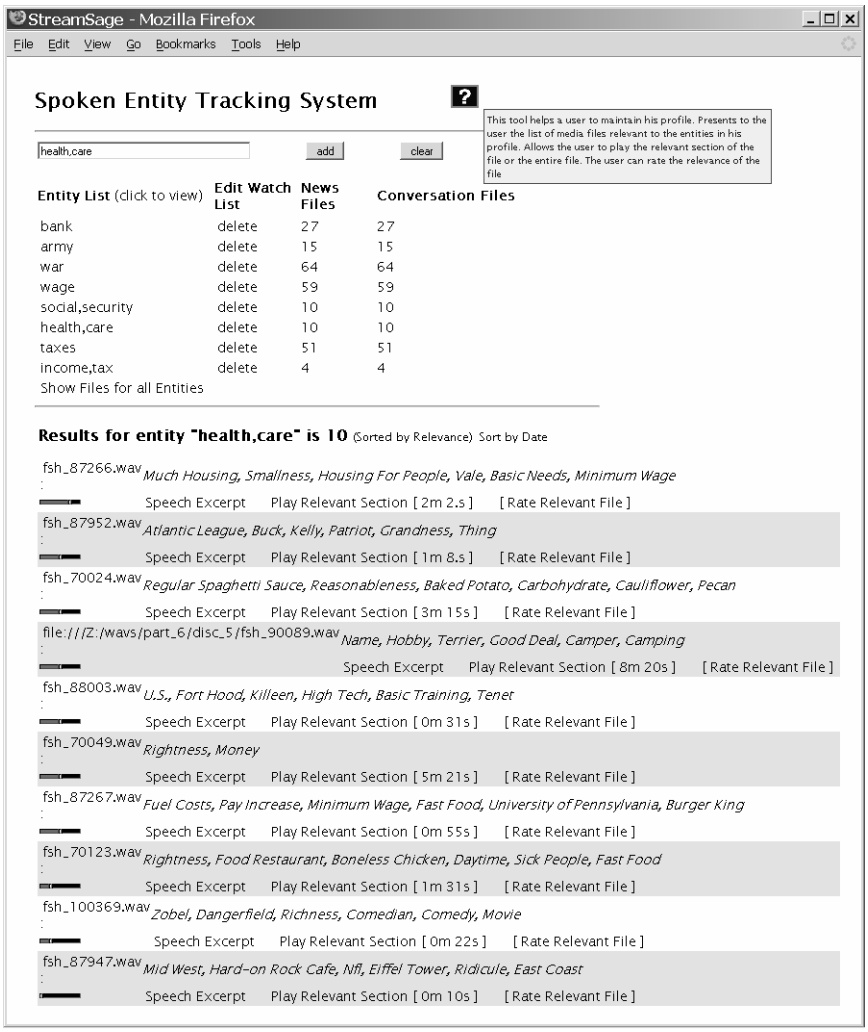
Upon entering the appropriate URL, the user sees the main screen of the demonstration analysts' platform. Using the most recent version of the stored profile, the system analyzes the available media files, and shows which files are relevant to the user's entities of interest.

For each relevant file, the system shows:

- How many minutes and seconds long the relevant segment is.
- A small graphical representation of where the relevant segment occurs in the source file.
- All the topic words for that segment. This lets the analyst gather, at a glance, what topic, related to the entity of interest, was discussed in that segment.

The platform also allows the user to view any selected segment (in an external window). The platform can also display the transcript for any segment.

Figure 4: the main screen of the analysts' platform.

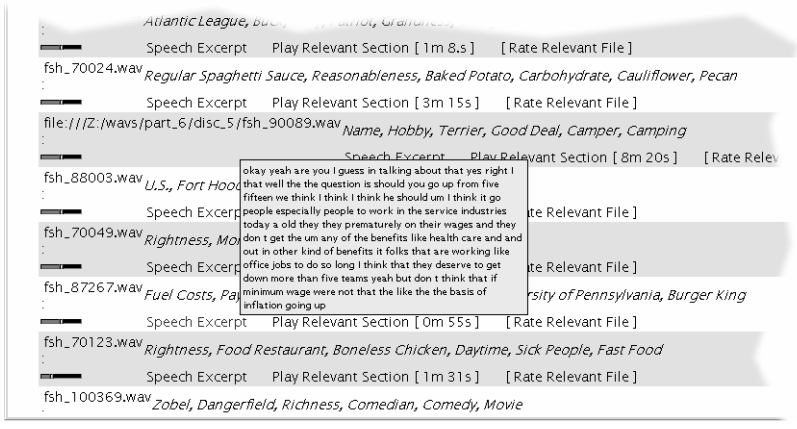


This profile includes ten entities of interest. The system is listing the segments relevant to one of them, "health care".

Note the ability to delete any entity from the profile, as well as the text field near the top for adding new entities to the profile.

Below "**Results for entity "health,care"**" you see the list of relevant media segments.

Figure 5: hovering over "Speech Excerpt", the analyst can view a transcript of that segment.



10. Bibliography

Alembic Workbench – Mitre - http://www.mitre.org/tech/alembic-workbench/alembic-workbench.html.

Allan J., J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pages 194-218, Lansdowne, VA, February 1998.

Appelt, D., & D. Martin, Named Entity Recognition in Speech: Approach and Results Using the TextPro System. *Proc. DARPA Broadcast News Workshop*, pp. 51-54, Herndon, VA, 1999.

Berger, A.L and Mittal, V.O (2000). *OCELOT:A system for summarizing web pages*. In *Proceedings of the 23rd Annual International ACM SIGIR*, Athens, Greece, pp 144-151.

Burger John D., Lynette Hirschman, and David D. Palmer, Named Entity Scoring for Speech Input, In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics* (ACL-COLING98), August 1998.

Chen, Langzhou, Jean-Luc Gauvain, Lori Lamel, Gilles Adda, and Martine Adda, Language Model Adaptation For Broadcast News Trancription. In *Proc. ISCA ITRW 2001 Adaptation Methods for Speech Recognition*, Sophia- Antipolis, Aug. 2001.

Daume III Hal, Abdessamad Echihabi, Daniel Marcu, Dragos Stefan Munteanu, and Radu Soricut (2002). GLEANS: A Generator of Logical Extracts and Abstracts for Nice Summaries. *Proceedings of the Document Understanding Conference (DUC-2002)*, Philadelphia, PA, July 11-12.

Furui, Sadaoki, Automatic speech recognition and its application to information extraction, *Proc. 37th meeting of ACL, Maryland*, U.S.A., pp.11-20, 1999.

Gaizauskas R. and K. Humphreys. Quantitative Evaluation of Coreference Algorithms in an Information Extraction System, Technical report CS-97-19, *Department of Computer Science, University of Sheffield*, 1997.

Gildea, Daniel, and Thomas Hofmann, Topic-Based Language Models Using EM. *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, 1999.

Hearst M. A., TexTiling: Segmenting text info multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33-64, 1997.

Ho, T.K, *The random subspace method for constructing decision forests*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 1998.

Hovy, E.H (2000). *Automated Text Summarization*. In R. Mitkov, editor, Oxford University Handbook of Computational Linguistics. Oxford Univ. Press.

Klavans, J.L. and Kan, M-Y. (1998). *Role of verbs in document analysis.* In proceedings of COLING/ACL 98.

Klavans, J.L. and Evelyne Tzoukermann, "Combining Corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons", in *The Machine Translation Journal*, Kluwer, 1996:10(3-4): 185-218.

Kubala, Francis, R. Schwartz, R. Stone, and R. Weischedel, Named entity extraction from speech, in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, (Lansdowne, VA), February 1998.

Miller David, Sean Boisen, Richard Schwartz, Rebecca Stone, Ralph Weischedel, Named Entity Extraction from Noisy Input: Speech and OCR, 1999.

Mooney, R.J and Cardie, C. *Symbolic Machine Learning for Natural Language Processing*. ACL'99 Tutorial, 1999.

McKeown, Kathleen, Regina Barzilay, Sasha Blaire-Goldensohn, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Barry Schiffman, Sergey Sigelman, "The Columbia Multi-Document Summarizer for Document Understanding Conference", in *Proc. of DUC*, 2002.

NIST 1999 Information Extraction - www.nist.gov/speech/tests/ie-er/er_99/er_99.htm

Przybocki Mark A., Jonathan G. Fiscus, John S. Garofolo, David S. Pallett, 1998 HUB-4 information Extraction Evaluation, *National Institute of Standards and Technology* (NIST), 1998.

Schiffman Barry, Ani Nenkova and Kathleen McKeown, Experiments in Multidocument Summarization, in *Proceedings of HLT 2002 Human Language Technology Conference*, San Diego, CA, 2002.

Sibley, Timothy, Audio & Video as a Flexible Information Source: Unlocking Potential through Sophisticated Language Analysis, *International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science, e-Medicine, and Mobile Technologies on the Internet, Italy*, January 2003.

Shriberg, E. & A. Stolcke, Prosody Modeling for Automatic Speech Understanding: An Overview of Recent Research at SRI. In M. Bacchiani, J. Hirschberg, D. Litman, & M. Ostendorf (eds.), *Proc. ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, pp. 13-16, Red Bank, NJ, 2001.

Shriberg, E., A. Stolcke, D. Hakkani-Tur, & G. Tur, Prosody-Based Automatic Segmentation of Speech into Sentences and Topics, *Speech Communication* 32(1-2), 127-154 (Special Issue on Accessing Information in Spoken Audio), 2000.

Srihari, Rohini, Cheng Niu, and Wei Li, A hybrid Approach for Named Entity and Sub-Type tagging, *1998.*

Srihari, Rohini and Wei Li, Information extraction supported question answering, in *Voorhees and Harman* [21], 1998.

Stolcke, A., E. Shriberg, D. Hakkani-Tur, G. Tur, Z. Rivlin, & K. Sonmez, Combining Words and Speech Prosody for Automatic Topic Segmentation, *Proc. DARPA Broadcast News Workshop*, pp. 61- 64, Herndon, VA, 1999.

Strzalkowski, T., Lin, F., Wang, J., and Perez-Carballo, J. Evaluating natural language processing techniques for information retrieval. In T. Strzalkowski, editor, Natural Language Information Retrieval. Kluwer, Boston, MA, 1999.

Tzoukermann, Evelyne, Smaranda Muresan, and Judith L. Klavans, GIST-IT: Combining Linguistic and Machine Learning Techniques for Email Summarization, *Human Language Technology and Knowledge Management (HLT/KM), ACL/EACL Conference*, Toulouse, France, 2001.

Ueda, Y., Oka M., Koyama T. and Miyauchi T (2000). Toward the "at-a-glance" summary: Phrase-representation summarization method. In *Proceedings of COLING* 2000.

Wacholder, N. *Simplex NPS sorted by head: a method for identifying significant topics within a document*, In Proceedings of the COLING-ACL Workshop on the Computational Treatment of Nominals, 1998.

Witten, I.H, Paynter, G.W., Frank E., Gutwin C. and Nevill-Manning, C.G. *KEA: Practical automatic key phrase extraction*. In Proceedings of DL'99, pp 254-256, 1999.

Wu, Jun, and Sanjeev Khudanpur, Building A Topic-Dependent Maximum Entropy Language Model for Very Large Corpora, *In Proceedings of ICASSP2002*, Orlando, USA, May 12-17, 2002.

Yamron, J., I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. A hidden Markov model approach to text segmentation and event tracking. In *Proc. ICASSP*, volume 1, pages 333-336, Seattle, WA, May 1998.